Statistique descriptive
Chapitre 7 : Régression linéaire simple
Moindre Carrés Ordinaire (MCO)

Mohamed BOUINCHA
m.bouincha@gmail.com
Institut National de Statistique et d'Économie Appliquée

Année universitaire : 2024/2025

INSEA

Sommaire

1. Rappel sur l'étude de la liaison entre deux variables quantitatives
2. Régression linéaire simple
3. Moindre Carrés Ordinaires (MCO)
4. Cas d'une relation non linéaire

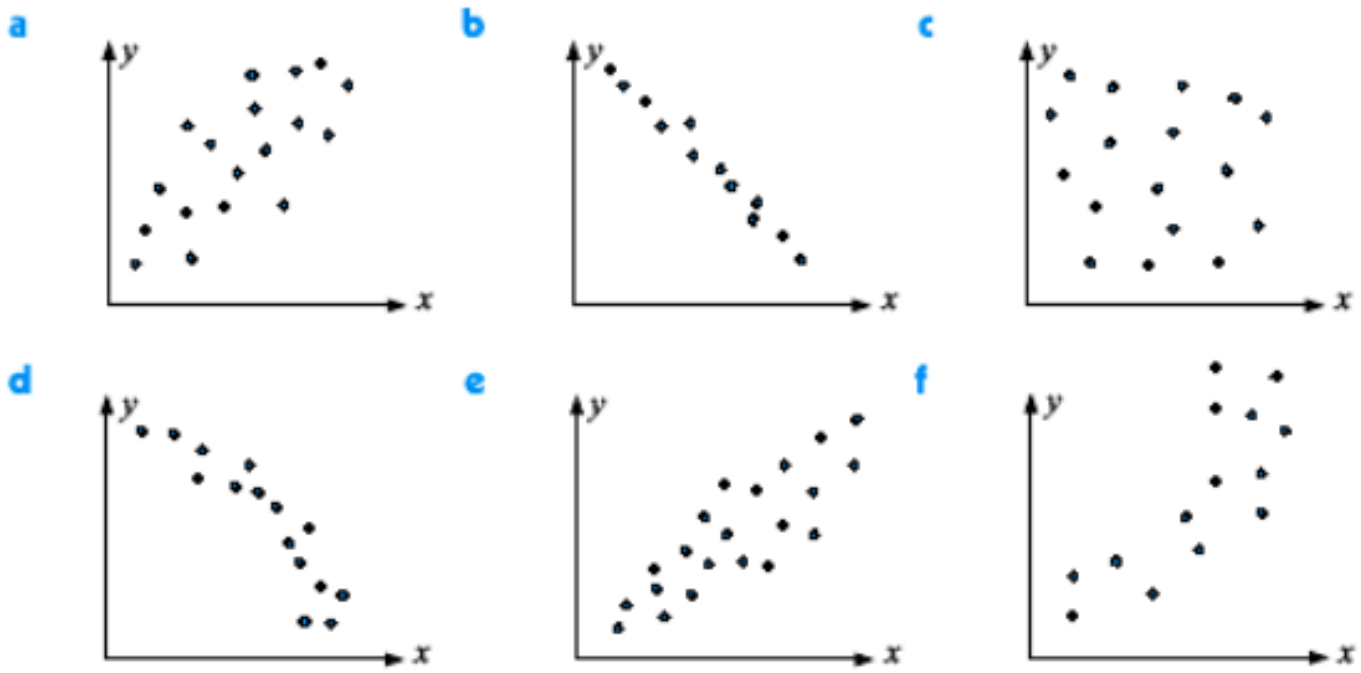
INSEA

Liaison entre deux variables quantitatives

- Pour étudier la liaison entre deux variables quantitatives, on peut utiliser :
  1. Le nuage des points ;
  2. Le coefficient de corrélation ;
  3. La régression linéaire simple ;

Nuage de points

- Comme déjà vu dans le chapitre précédent, le nuage de point une une idée graphique sur la nature de liaison entre deux variables ;
- La figue suivante présente plusieurs formes de nuage de points ;
- Quoi que le nuage de points présente une visualisation de la liaison entre les deux variable, il ne permet pas de statuer avec exactitude sur la nature de la relation

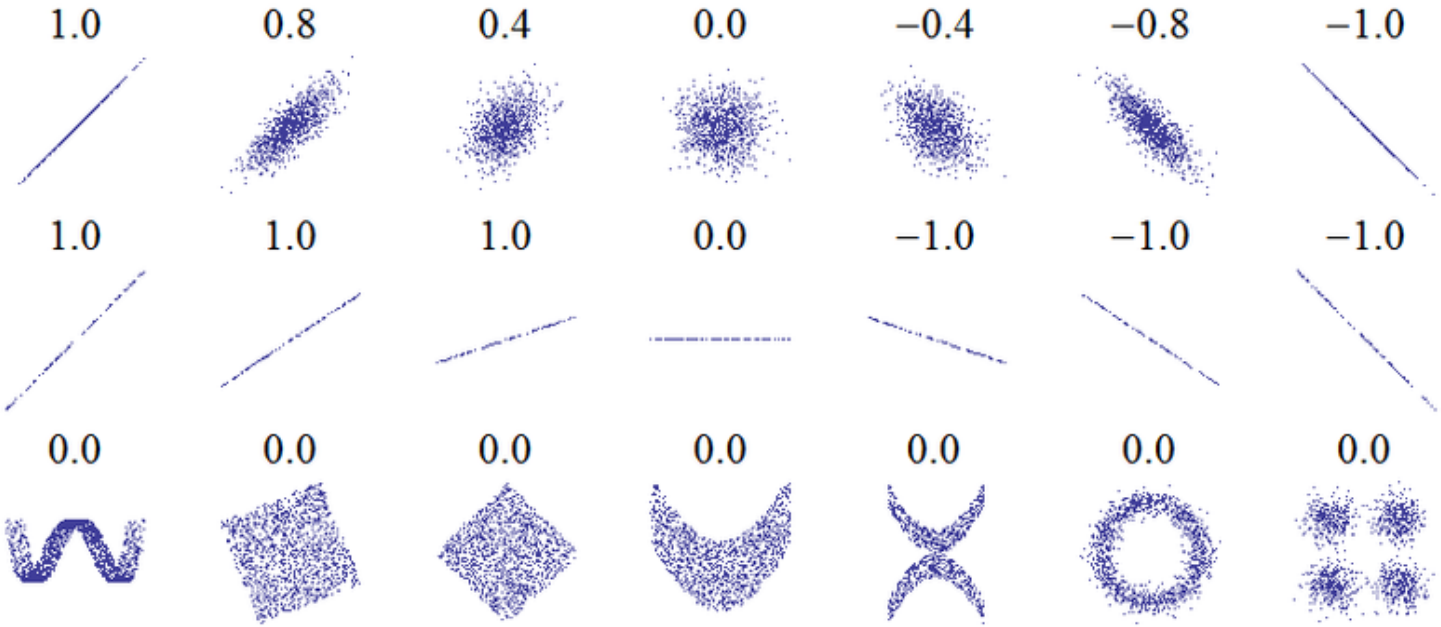


Le coefficient de corrélation

- Comme nous avons vu dans le chapitre précédent, le coefficient de corrélation donne une mesure standardisée de la liaison entre deux variables ;
- Soit l'exemple suivant :
  1. Y : Le revenu du ménage ;
  2. X : L'âge du chef du ménage ;
  3. D'après les données, nous avons trouvé un coefficient de corrélation égale à r=0,53 ;
  4. Forte corrélation : Il semble que le revenu du ménage augmente avec l'âge du patient ;
  5. Mais, peut-on conclure que le coefficient de corrélation linéaire ρ de la population est significativement différent de 0 ?
  6. Pour répondre à cette question, il faut mener un test d'hypothèse statistique (un prochain cours) ;

Limites du coefficient de corrélation

- Attention aux conclusions non-valides pour une relation non-linéaire ;
- Plusieurs relations possibles pour un même coefficient de corrélation, comme montré dans la figure ci-dessous ;
- Aucune quantification de la relation.



Régression linéaire simple

- La régression linéaire (simple et multiple) permet de faire :
  1. Prédiction : Elle permet de prédire la valeur d'une variable en fonction d'autres variables connues. Par exemple, prédire le prix d'une maison en fonction de sa superficie, de son emplacement, etc.
  2. Analyse des relations : Elle aide à comprendre comment les variables sont liées les unes aux autres. Par exemple, déterminer l'influence des heures d'étude sur les résultats aux examens.
  3. Contrôle et ajustement : Elle est utilisée pour contrôler et ajuster les variables indépendantes afin de comprendre comment ces ajustements pourraient affecter la variable dépendante.
  4. Identification des variables importantes : Elle peut aider à identifier les variables qui ont le plus d'impact sur la variable cible, en éliminant celles qui ont peu ou pas d'influence.
  5. Évaluation des tendances et des prévisions : Elle permet de modéliser des tendances historiques et de faire des prévisions basées sur ces modèles.

INSEA

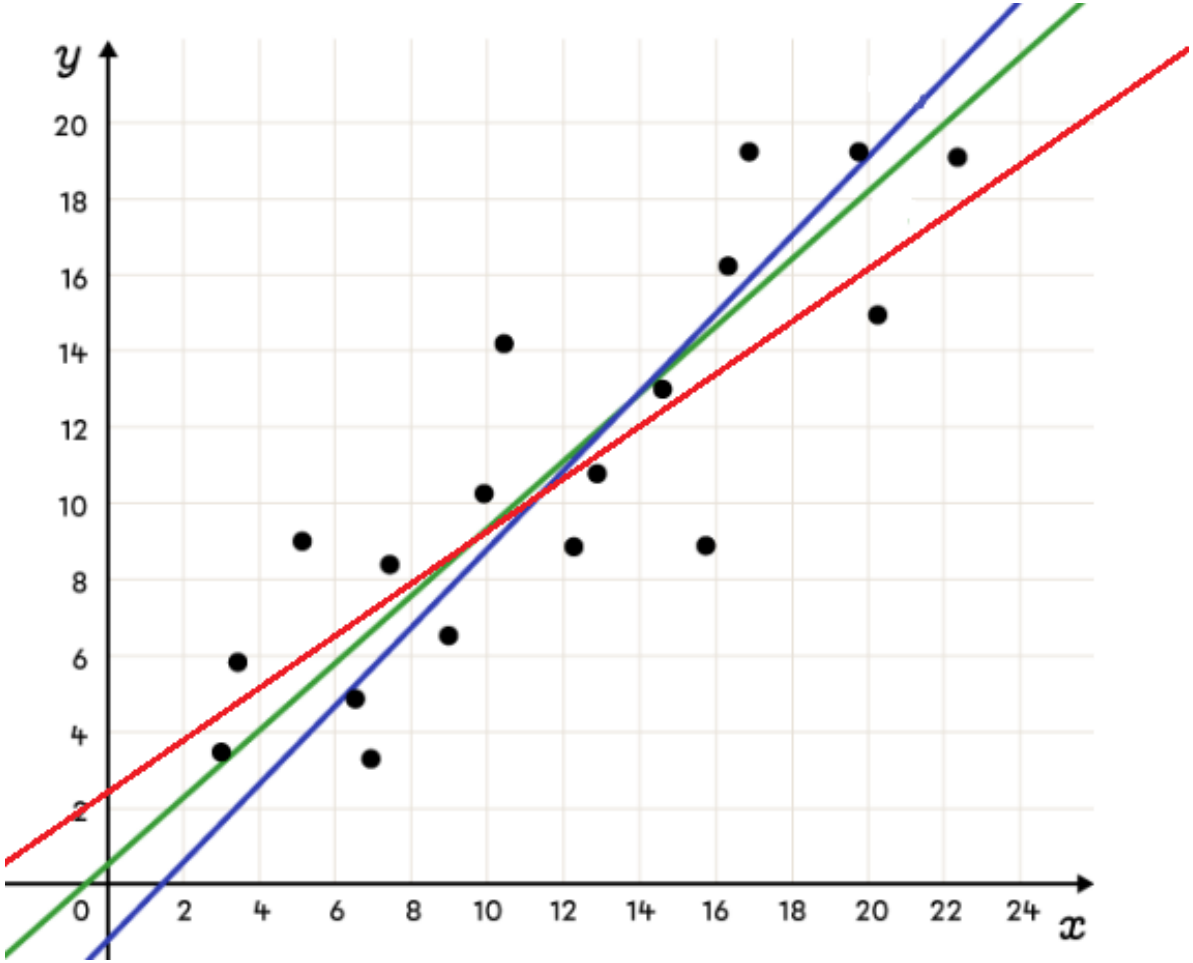
Régression linéaire simple

- La régression linéaire simple est une méthode statistique qui cherche à établir une relation linéaire entre une variable indépendante (explicatrice) et une variable dépendante (cible) ;
- Le principe de la régression linéaire simple repose sur l'idée de modéliser cette relation linéaire sous la forme d'une droite ;
- On suppose qu'il existe une relation linéaire entre la variable indépendante (X) et la variable dépendante (Y) ;
- L'objectif est de trouver la droite qui minimise la somme des carrés des erreurs (différences entre les valeurs prédites et les valeurs réelles) ;
- Les valeurs réelles sont celles dont nous disposons, elles proviennent souvent de l'historique des variables ;
- Les valeurs prédites sont celle calculées en utilisant la relation de la ligne de régression.

INSEA

Régression linéaire simple : le modèle

- La régression linéaire simple permet de choisir la ligne qui donne une meilleur représentation du nuage de points



- Quelle est la couleur de la droite qui vous apparaît proche du maximum de points ? Comment faire le choix ?

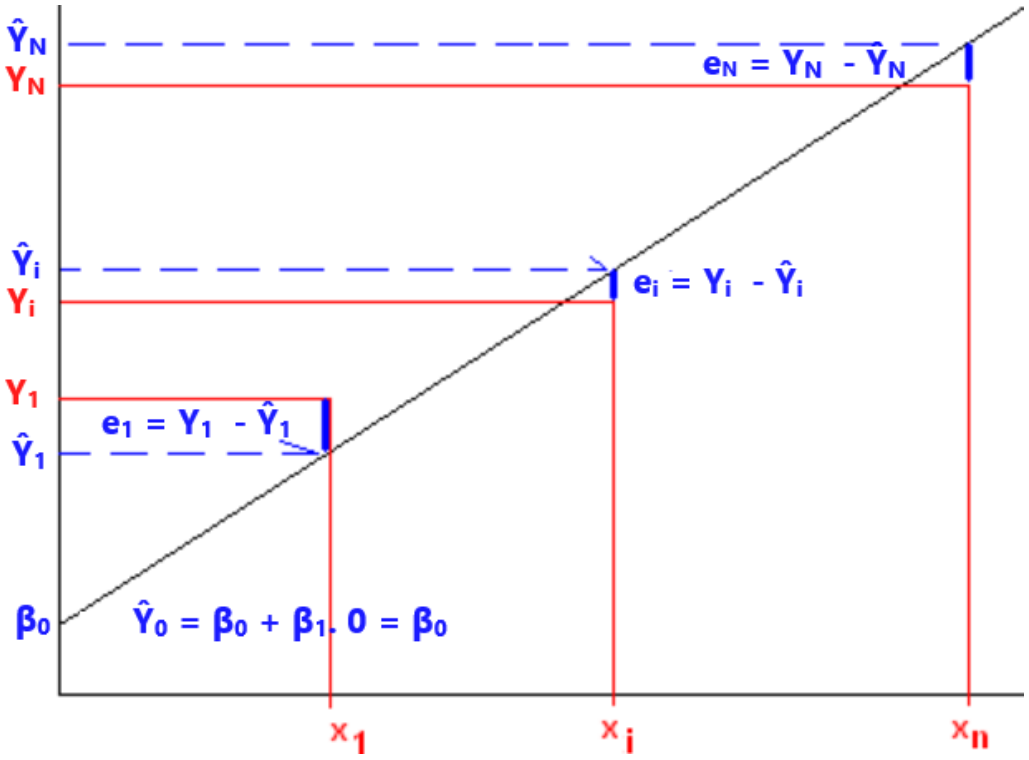
Régression linéaire simple : le modèle

- la fonction de la ligne à estimer en utilisant la régression linéaire simple se présente comme suit :

$$Y = \beta_0 + \beta_1.X_i$$

- On a les points (X_i, Y_i) pour i=1,...,N
- Pour chaque X_i, on associe une ordonnée $\hat{Y}_i = \beta_0 + \beta_1.X$ qui représente la valeur de Y ajustée par le modèle linéaire $Y = \beta_0 + \beta_1.X$;
- Un écart e_i représente la différente entre la valeur réelle de Y pour l'individu i et la valeur $\hat{Y}_i$ calculée en utilisant la ligne de régression ;
- Souvent, on utilise le terme ϵ_i au lieu de e_i ;
- Donc :
  $e_i = \epsilon_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1.X_i)$
- L'estimation de la droite de la régression se fait souvent par la méthode des Moindre Carrés Ordinaires (MCO) ;
- Mais, on peut utiliser aussi la méthode de Lagrangien (optimisation non linéaire) ;

Régression linéaire simple : le modèle



- β_0 représente la valeur de $\hat{Y}$ pour $X_i = 0$. Il représente l'intersection entre la droite de régression et l'axe des ordonnées ;
- β_1 représente la pente de la droite de la droite de régression. Il représente la quantité de Y due à l'augmentation de X par une unité ;

Méthode des Moindre Carrés Ordinaires (MCO)

- Le problème de la méthode des Moindre Carrés Ordinaires (MCO) consiste à chercher les paramètres de la droite qui minimise le carré des écarts ;
- **Min** $Q(\beta_0, \beta_1) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N}(Y_i - (\beta_0 + \beta_1.X_i))^2$
- On dérive $Q(\beta_0, \beta_1)$ par rapport à chacun des paramètres à estimer β_0 et β_1 ;
- On minimise le carré des résidus car la somme des écarts non pris au carrés est nulle. Autrement :
  $\sum_{i=1}^{N} e_i = \sum_{i=1}^{N}(Y_i - \hat{Y}_i) = 0$. Les écarts positifs compensent ceux négatifs ;
- On égalise les dérivées premières à 0 :
  1. $\frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_0} = 0 \Rightarrow \widehat{\beta_0}$
  2. $\frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_1} = 0 \Rightarrow \widehat{\beta_1}$

INSEA

La droite de régression

- La méthode des M.C.O donne les paramètres suivants de la droite de régression :
  1. $\beta_1 = \frac{Cov(X,Y)}{V(X)} = \frac{\sum_{i=1}^{N}(X_i-\bar{X}).(Y_i-\bar{Y})}{\sum_{i=1}^{N}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{N} X_i.Y_i - N.\bar{X}.\bar{Y}}{\sum_{i=1}^{N} X_i^2 - N.\bar{X}^2}$
  2. $\beta_0 = \bar{Y} - \beta_1.\bar{X}$
- La droite de régression passe nécessairement par les moyennes $\bar{X}$ et $\bar{Y}$;

La qualité de la régression : le coefficient de détermination

- Pour apprécier la qualité de l'ajustement de la droite des M.C.O, on calcule **le coefficient de détermination** R^2 ;
- Le coefficient R^2 est une mesure utilisée pour évaluer à quel point le modèle linéaire ajusté s'ajuste aux données réelles ou observées ;
- Il représente la proportion de la variance de la variable dépendante (Y) qui est expliquée par la variable indépendante (X) ;
- En termes simples, R^2 mesure la proximité des points de données par rapport à la droite de régression ;
- R^2 est toujours entre 0 et 1 : Plus il est proche de 1, plus la droite s'ajuste bien aux données. S'il est proche de 0, la droite n'explique pas bien la variance des données ;
- **Le coefficient R^2 n'est pas le carré du coefficient $\rho_{X,Y}$ de corrélation entre X et Y !**

INSEA

Calcul du coefficient R^2

- On a :
  1. $\hat{Y}_i = \beta_0 + \beta_1.X_i$: les observations de $Y_{ajustée}$ que nous avons calculé à partir de la droite de régression ;
  2. $e_i = Y_i - \hat{Y}_i$
  3. $\underbrace{Y_i}_{Total} = \underbrace{\beta_0 + \beta_1.X_i}_{droite} + \underbrace{e_i}_{erreurs}$
- Pour calculer le coefficient R^2, on a :
  1. **Var(Y) =** $\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N}$: Représente la variance totale de la variable Y ;
  2. **Var**$(\hat{Y}) = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{\hat{Y}_i})^2}{N}$: Représente la variance des observations de $\hat{Y}$ estimées par la droite de régression ;
  3. **Var**$(e_i) = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}$: Représente la variance des erreurs ;
  4. $\underbrace{V(Y)}_{\textit{Variance totale}} = \underbrace{V(\hat{Y})}_{\textit{Variance expliquée}} + \underbrace{V(e)}_{\textit{Variance résiduelle}}$
- Donc :

$$R^2 = \frac{V(\hat{Y})}{V(Y)} = 1 - \frac{V(e)}{V(Y)} = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

INSEA

M.C.O : Exemple d'application

- Soient les données sur la consommation des médicaments et la population :

Pays (i)	Cons(Y)	Pop.(X)
1	327,4	17,3
2	179,5	7,7
3	279,4	10,4
4	139,1	5,1
5	92,5	5
6	926,7	56,9
7	2186,3	79,7
8	523,9	57,8
9	935,9	124
10	444,2	15,1
11	119,7	4,3
12	300,7	39
13	201,9	8,7
14	194,7	6,9
15	1592,9	57,7
16	5142,2	252,7

1. Calculer le coefficient de corrélation $r_{X,Y}$
2. Calculer des paramètres de la droite de régression β_0 et β_1
3. Calculer le coefficient de détermination R^2 et commenter la qualité de l'ajustement ;

M.C.O : Corrigé de l'exemple d'application

i	X	Y	$X_i.Y_i$	$(X_i)^2$	$(Y_i)^2$
1	17	327	5 664	299	107 191
2	8	180	1 382	59	32 220
3	10	279	2 906	108	78 064
4	5	139	709	26	19 349
5	5	93	463	25	8 556
6	57	927	52 729	3 238	858 773
7	80	2 186	174 248	6 352	4 779 908
8	58	524	30 281	3 341	274 471
9	124	936	116 052	15 376	875 909
10	15	444	6 707	228	197 314
11	4	120	515	18	14 328
12	39	301	11 727	1 521	90 420
13	9	202	1 757	76	40 764
14	7	195	1 343	48	37 908
15	58	1 593	91 910	3 329	2 537 330
16	253	5 142	1 299 434	63 857	26 442 221
Total	**748**	**13 587**	**1 797 828**	**97 902**	**36 394 726**

INSEA

M.C.O : Corrigé de l'exemple d'application

- $\bar{X} = \frac{748}{16} = 46.77$
- $\bar{Y} = \frac{13587}{16} = 849.19$
- $\rho_{X,Y} = \frac{N.\sum_{i=1}^{N} X_i.Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{\sqrt{N.\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2}.\sqrt{N.\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2}} =$
  $\frac{16x1797828-748x13587}{\sqrt{(16x97902)-(748)^2}x\sqrt{(16x36394726)-(13587)^2}} = 0,93$: Une forte liaison positive ;
- Les paramètres de la droite de régression :
  1. $\beta_0 = \bar{Y} - \beta_1.\bar{X} = 849,19 - (18,48x46,77) = -15,03$
  2. $\beta_1 = \frac{\sum_{i=1}^{N} X_i.Y_i - N.\bar{X}.\bar{Y}}{\sum_{i=1}^{N} X_i^2 - N.\bar{X}^2} = \frac{1797828-16x46.77x849.19}{97902-16x(46,77)^2} = 18,48$: **La consommation des pays augmente de 18.48 unitées avec l'augmentation de la population d'une unité ;**
- Pour calculer le coefficient de détermination R^2, on doit d'abord calculer les $\hat{Y}_i$;

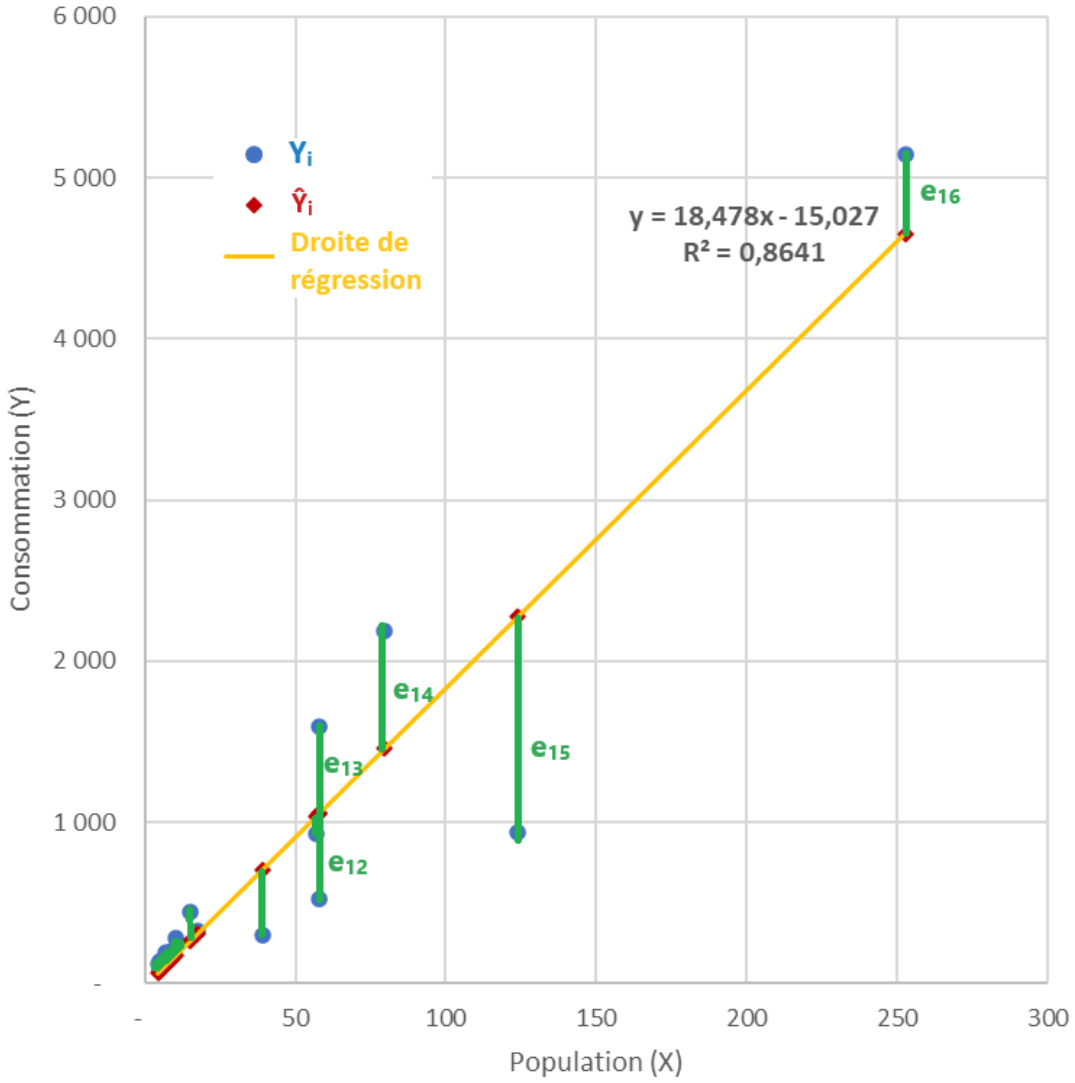
INSEA

M.C.O : Corrigé de l'exemple d'application

X	Y	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
17	327	- 522	272 262	305	23	518
8	180	- 670	448 481	127	52	2 729
10	279	- 570	324 658	177	102	10 455
5	139	- 710	504 224	79	60	3 586
5	93	- 757	572 576	77	15	229
57	927	78	6 008	1 036	- 110	12 034
80	2 186	1 337	1 787 870	1 458	729	530 848
58	524	- 325	105 812	1 053	- 529	279 977
124	936	87	7 519	2 276	- 1 340	1 796 679
15	444	- 405	164 015	264	180	32 473
4	120	- 729	532 152	64	55	3 055
39	301	- 548	300 839	706	- 405	163 971
9	202	- 647	418 981	146	56	3 154
7	195	- 654	428 354	112	82	6 761
58	1 593	744	553 108	1 051	542	293 460
253	5 142	4 293	18 429 956	4 654	488	237 869
748	**13 587**	**0**	**24 856 816**	**13 587**	**0**	**3 377 799**

M.C.O : Corrigé de l'exemple d'application

- $R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{3377799}{24856816} = 1 - 0,1359 = 0,8641$;



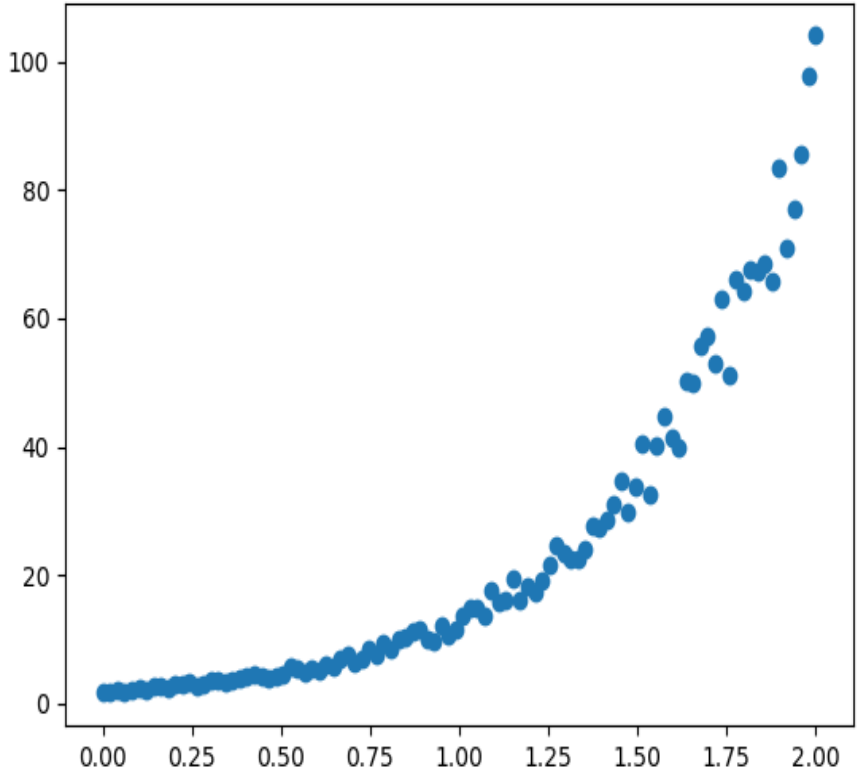
Cas d'une relation non linéaire

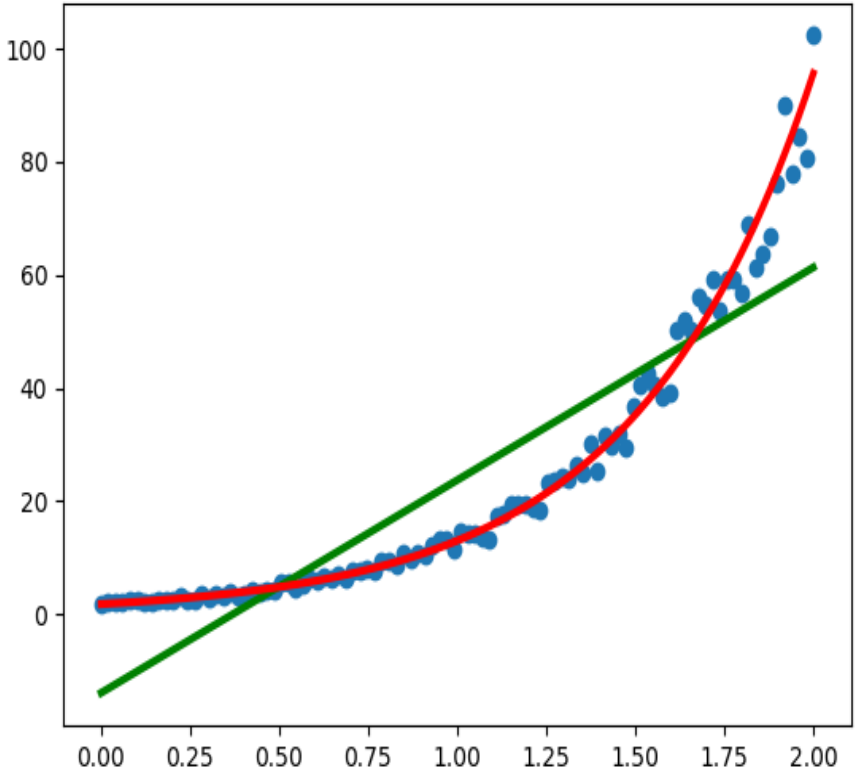
- L'utilisation de la méthode des M.C.O se base sur l'hypothèse de linéarité entre X et Y ;
- En cas de non linéarité, il faut faire une transformation des variables avant d'utiliser les M.C.O ;
- Nous allons voir les trois cas suivants :
  1. La relation exponentiel ;
  2. La relation à élasticité constante ;
  3. La relation logistique ;
- **Attention : La lecture des paramètres de la droite de régression change si on introduit le logarithme sur une des variables du modèle ou sur les deux ou tout simplement si la relation entre les deux variables X et Y n'est linéaire.**

INSEA

Relation non linéaire : Schéma exponentiel

- Si le nuage de points présente une forme exponentiel, on utilise une équation de relation exponentielle ;
- Si on utilise un modèle linéaire, le coefficient de détermination va être faible comme montré dans la figure de droite ;



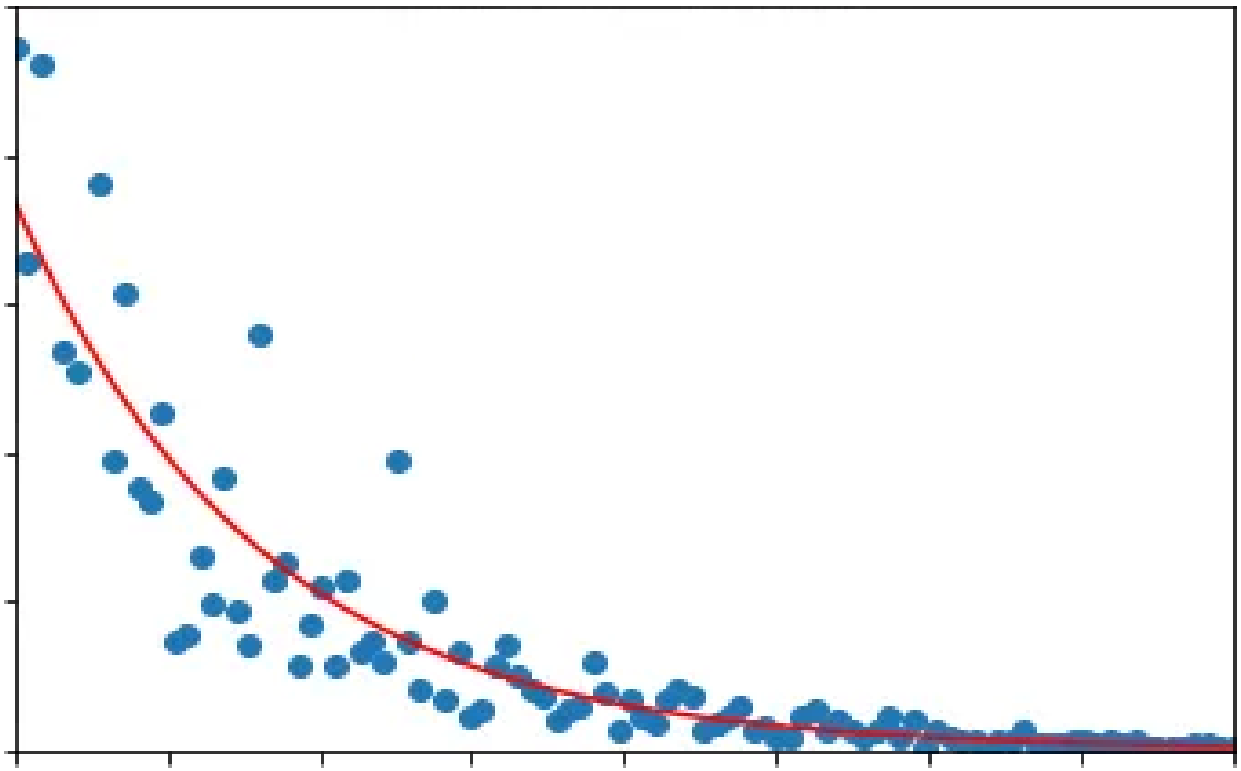


Relation non linéaire : Schéma exponentiel

- L'équation de la relation exponentielle est comme suit :
- $Z = k.e^{a.X}$
- Si on introduit le logarithme des deux cotés, on aura :
- $ln(Z) = ln(k) + a.ln(X)$, la fonction est maintenant sous la forme linéaire $Y = \beta_0 + \beta_1.X$. Avec :
  1. Y = ln(Z) ;
  2. X = X (inchangée)
  3. β_1 = a
  4. β_0 = ln(k)
- $\beta_0 = \overline{ln(Z)} - \beta_1.\bar{X}$
- $\beta_1 = \frac{Cov(X,ln(Z))}{V(X)} = \frac{\overline{X.ln(Z)} - \bar{X}.\overline{ln(Z)}}{\overline{X^2} - (X)^2}$
- Interprétation de β_1 : La variation **relative** de Y suite à la variation de X d'une unité ;

Relation non linéaire : Schéma à élasticité constante

- Si le nuage de points présente une forme d'une relation décroissante entre X et Y, mais avec un impact de X sur Y différent selon le niveau de X on utilise une équation de nature à élasticité constante ;
- Si on utilise un modèle linéaire, le coefficient de détermination va être faible ;
- Un exemple de la théorie économique est la relation entre la propension marginale à consommer et le revenu ;

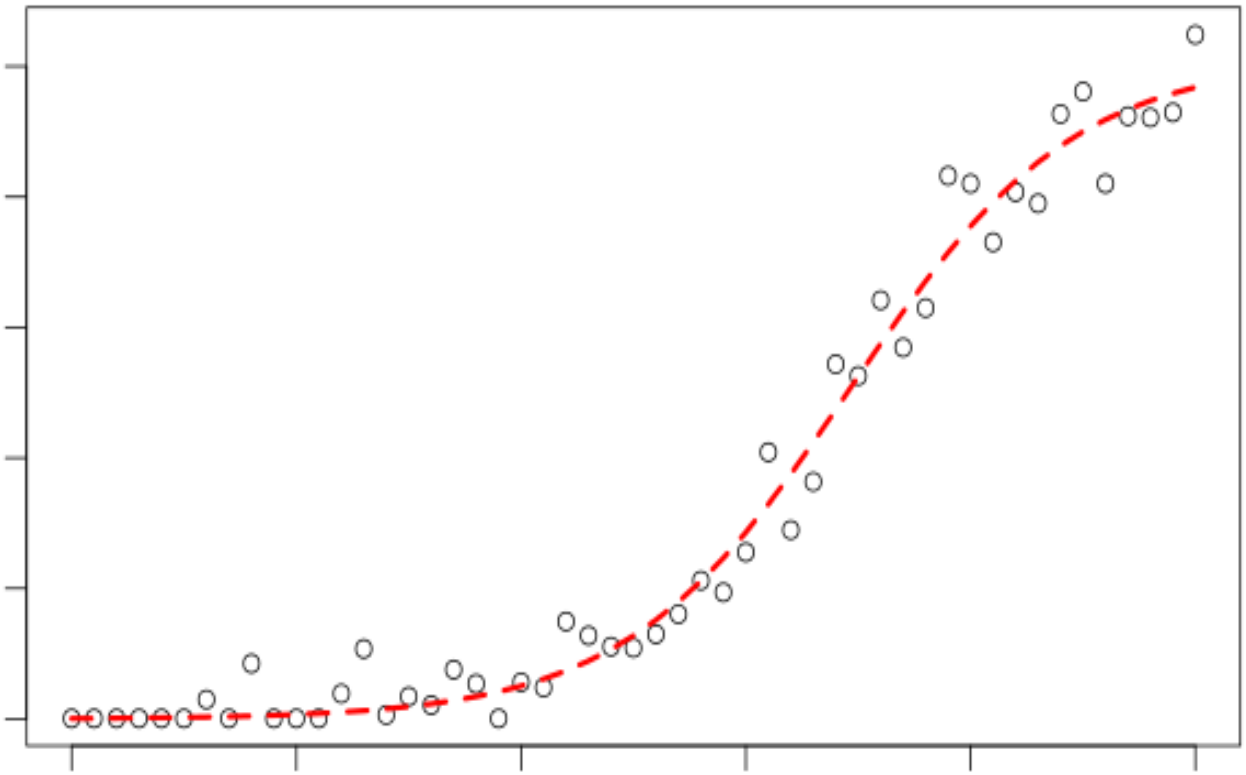


Relation non linéaire : Schéma à élasticité constante

- L'équation de la relation exponentielle est comme suit :
- $Z = k.U^a$
- Si on introduit le logarithme des deux cotés, on aura :
- $ln(Z) = a.ln(U) + ln(K)$ on a fonction est maintenant sous la forme linéaire $Y = \beta_0 + a.X$. Avec :
  1. Y = ln(Z) ;
  2. X = ln(U)
  3. β_1 = a
  4. β_0 = ln(k)
- $\beta_0 = \overline{ln(Z)} - \beta_1.\overline{ln(U)}$
- $\beta_1 = \frac{Cov(ln(U),ln(Z))}{V(ln(U))} = \frac{\overline{ln(U).ln(Z)} - \overline{ln(U)}.\overline{ln(Z)}}{\overline{[ln(U)]^2} - \overline{[ln(U)]}^2}$
- Interprétation de β_1 : La variation **relative** de Y suite à une variation **relative** de X ;

Relation non linéaire : Schéma logistique

- Si le nuage de points présente une forme d'une relation croissante entre X et Y, mais Y presque stagne lorsque X atteint un certain niveau, on utilise une équation de nature à élasticité constante ;
- Si on utilise un modèle linéaire, le coefficient de détermination va être faible ;



Relation non linéaire : Schéma logistique

- L'équation de la relation exponentielle est comme suit :
- $Z = \frac{1}{1+k.e^{-a.X}}$
- Pour linéariser la relation on suit les étapes suivantes :
  1. $\frac{1}{Z} = 1 + k.e^{-a.X}$
  2. $\frac{1}{Z} - 1 = k.e^{-a.X}$
  3. $\frac{1-Z}{Z} = k.e^{-a.X}$
  4. $\frac{1-Z}{Z} = \frac{e^{a.X}}{k}$
  5. $ln\left(\frac{1-Z}{Z}\right) = a.X - ln(k)$ la fonction est maintenant sous la forme linéaire $Y = \beta_0 + a.X$. Avec :
  1. $Y = ln\left(\frac{1-Z}{Z}\right)$;
  2. X = X (inchangée)
  3. β_1 = a
  4. β_0 = - ln(k)
- $\beta_0 = \overline{ln\left(\frac{1-Z}{Z}\right)} - \beta_1.\bar{X}$
- $\beta_1 = \frac{Cov(X,Y)}{V(X)} = \frac{\overline{X.ln\left(\frac{1-Z}{Z}\right)} - \bar{X}.\overline{ln\left(\frac{1-Z}{Z}\right)}}{V(X)}$
- Interprétation de β_1 : La variation **relative** de $Y = \frac{1-Z}{Z}$ suite à une variation de X d'une unité ;

Questions ?